Gap Based Inverse Sampling

Bardia Panahbehagh
Department of Mathematics, Faculty of Mathematics and computer science,
Kharazmi University, Tehran, Iran, panahbehagh@khu.ac.ir

Jennifer Brown
School of Mathematics and Statistics, University of Canterbury,
Christchurch, New Zealand

July 12, 2016

Abstract

We present a new inverse sampling design for surveys of rare events, Gap Based Inverse Sampling. In the design, sampling stops if, after a predetermined interval, or gap, no new rare events are found. The length of the gap that follows after finding a rare event is used as a way of limiting sample effort. We present stopping rules using decisions based on the gap length, the total number of rare events found, and a fixed upper limit of survey effort. We illustrate the use of the design with stratified sampling of two biological populations. The design uses the intuitive behaviour of a field biologist in stratified sampling, where if in a stratum nothing is found after a long search, the field surveyor would like to consider the stratum is empty and stop searching. Our design has appeal for surveying rare events (for example a rare species) with stratified sampling where there are likely to be some completely empty strata.

Keywords and phrases: Inverse Sampling, Sequential Sampling, Murthy Estimator, Rare Events.

1 Introduction

Inverse sampling is a form of adaptive sampling, where units are selected until a predefined number of events have been detected. For surveys of rare events, for example, surveys of a rare plant, or surveying for extreme environmental conditions, inverse sampling can be used as a way of limiting the amount of survey effort that is used. In these two examples, surveying would continue until the desired number of plants, or sample units where the plant occurs, are found; or the extreme environmental condition has been detected a predefined number of times. One of the first inverse sampling designs was introduced Haldane (1945) who used inverse sampling to estimate the frequency of a rare event. Finney (1949) proposed an unbiased estimator of the variance. More recently Chistman and Lan (2001) considered inverse sampling with, and without, replacement when the selection of units is with equal probability in each draw. An unbiased estimator of the population total and its variance were provided but an unbiased estimator of the variance was not given. In their work, Christman and Lan considered some inverse sampling designs that use stopping rules based on the number of rare units observed in the sample. Two of these rules were:

(i) Select units one at a time until a predetermined number of rare units, say k, in the sample is obtained.

(ii) First select an initial simple random sample of size n_1. If the number of rare units is greater than or equal to k stop sampling, otherwise continue sampling until k rare units are observed.

Salehi and Seber (2004) introduced a design based on the second of form proposed by Chistman and Lan (2001), but with a new stopping rule. The new stopping rule terminates sampling when effort reaches some predefined limit.

In this paper we extend the work of Salehi and Seber (2004) by adding a new condition to the design to improve its practical application. In our new design, Gap Based Inverse Sampling (GBIS), selection of units stops if no new rare event is found after a pre-specified gap in the sampling effort. This gap can be defined in different ways, for example, a sequence of selected units that fail to detect the rare event immediately following selection of a unit that does contain the rare event. One practical, and appealing, application of the new design is in stratified sampling for populations where there are likely to be a number of empty strata. In the proposed design where the interest is in searching for a rare event, the sample effort within empty (or near empty) strata will be redirected to other more fruitful strata.

2 Notation and sampling design

Following Salehi and Seber (2004), given a population $U = \{u_1, u_2, ..., u_N\}$ of N units, where N is known, let y_i denotes the y-value associated with u_i, for $i = 1, 2, ..., N$. The population is divided into two subpopulations according to whether the y-values satisfy some condition C, for example $C = \{y_i > c\}$, where c is a constant. The two subpopulations is indicated by $U_M = \{u : y_i \in C, i = 1, 2, ..., N\}$ and $U_{N-M} = \{u : y_i \notin C, i = 1, 2, ..., N\}$, where $M = |U_M|$ is the unknown number of units, or cardinality, of U_M. We suppose that the subpopulation to which a unit belongs is not known until the unit is sampled. The final sample is denoted by s that we partitioned into $s = s_C \cup s_{C'}$ where $s_C = s \cap U_M$ with size n_C and $s_{C'} = s \cap U_{N-M}$ with size $n_{C'}$.

Gap Based Inverse Sampling (GBIS) is designed as the following. Select a simple random sample of size n_1 and stop further sampling if at least k units from U_M are selected. Otherwise, we sequentially continue sampling until either exactly k units from U_M are selected, or we have faced a gap of size g, or n_3 units are selected in total, where $n_1 \leq n_3$.

We now consider the design for estimating the population mean $\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i$.

3 Murthy Estimator for the Designs

Unbiased estimators for the design, using Murthy estimator (based on Salehi and Seber 2001) will be constructed in the next theorem.

Theorem 1 *In GBIS an unbiased estimator for μ_y is*

$$\hat{\mu}_y = \begin{cases} \frac{\Phi(1,g)}{\Phi(0,g)}\bar{y}_{sc} + \frac{\Phi(0,g+1)}{\Phi(0,g)}\bar{y}_{sc'} & 1 < r < k \\ \frac{\Phi(1,0)}{\Phi(0,0)}\bar{y}_{sc} + \frac{\Phi(0,1)}{\Phi(0,0)}\bar{y}_{sc'} & r > 1, r = k \\ \frac{\Phi(0,0)}{\Phi(-1,0)}\bar{y}_{sc} + \frac{\Phi(-1,1)}{\Phi(-1,0)}\bar{y}_{sc'} & r < k, n = N \ \ or \ \ r > 1, g_n < g, n = n_3 \\ y_1 & r = 1, y_1 \in C \\ \bar{y}_{sc'} & r = 1, y_1 \in C' \\ \bar{y}_s & r = 0 \ \ or \ \ n = n_1 \end{cases}$$

where

$$\Phi(h,m) = \sum_{j=0}^{n_c-h}(-1)^j \binom{n_c-h}{j}\binom{n_{c'}-m+n_c-h-1-jg}{n_c-h-1},$$

r is number of rare units in the final sample and g_n is the gap for the last unit satisfying C.

For the proof of the theorem see appendix 1.

4 Two Simulation Studies on Real Data

4.1 Buttercup Population

To evaluate our proposed gap based design we used a case study of a rare buttercup population (Brown 2010). The data are from a study in 1998 of a buttercup found within the Lance McCaskill Nature Reserve in the South Island of New Zealand. The Castle Hill buttercup (*Ranunculus crithmifolius* sub. *paucifolius*) is one of New Zealands rarest plants. Locations of buttercup plants observed were mapped within 10×10m quadrants. The counts of plants within 300 quadrants are shown in figure 1, and with the area divided into 12 equal-sized strata. We used stratified sampling to compare GBIS with General Inverse Sampling (GIS), the design proposed by Salehi and Seber (2004) and with the conventional Inverse Sampling (CIS).

In this example the rare event is when at least one buttercup is found in a quadrant. We define the gap as being the sequence of counts of quadrants visited where no buttercups are found (in other words the rare event is not found).

We compared the designs by simulating samples with different set-up options by varying n_1, n_3, k and g. Here n_1, n_3, k and g represent the within-stratum values. The within-stratum population mean was estimated for GBIS using the Murthy estimators shown above and for CIS and GIS using their conventional Murthy estimators.

The conventional estimator of stratified sampling was then used to estimate the overall population mean. We used a Monte Carlo approach and for each of the three designs; CIS, GIS, GBIS, simulated 10000 samples with the different values for n_1, n_3, k and g.

To evaluate and compare the efficiency of the designs, we considered CIS as the base design and used this to compare the efficiency of GBIS and GIS. The inverse sampling designs will often results in differences among the final sample sizes even with the same set-up options. We followed the methods of Wang et al. (2004) and defined relative efficiency to include sample effort:

$$RE_{.} = \frac{Var(CIS)}{Var(.)} \times \frac{E(n_{\mathrm{CIS}})}{E(n_{.})}$$

where $Var(CIS)$ and $E(n_{\mathrm{CIS}})$ are the variance of the Murthy estimator and the expectation of the final sample size in CIS respectively, and . stands for GIS and GBIS.

The largest final sample size among all set-up options was for CIS (Table 1). The final sample size on average was almost twice the size for GBIS. The smallest final sample size was with GBIS. GBIS presents three alternatives for sampling to cease within a stratum: either because there is a gap of sufficient survey-effort length, or sufficient rare events are detected, or because the maximum sample size has been reached. This broader class of stopping rules has the effect of limiting the final sample size.
Selection of the set-up values for n_1, n_3, k and g will affect the final sample size, with larger final sample sizes with increases in all of these values. Clearly, as the initial sample size, the desired number of rare events, the length of gap that is needed to stop sampling, and, upper limit to the sample size, all increase, the final sample size increases accordingly.

The new gap based sampling design, GBIS, was also superior in terms of sample efficiency. The relatively efficiency on average was 3.13 for GBIS (Table 1). In contrast GIS relative efficiency on average was 2.55. The relative efficiencies were generally highest when sampling was set up to have the smaller gap lengths and smaller numbers of rare events that needed to be detected for sampling to stop. For example, with small g or small k, the relative efficiency of all designs was improved compared with large g or large k. The size of the initial sample for GBIS had an effect on the relative efficiency, with improvements in efficiency when the initial sample size, n_1, was larger. However, efficiency improved when the sample effort within strata was controlled by a smaller value of n_3. For the designs other than with the smallest value of g and k, the relative efficiency of GBIS was higher with the smaller value of n_3.

The comparative study was repeated with different scales of stratification. With more coarse stratification, that is, with six strata rather than 12, the largest final sample sizes among all set-up options again were for CIS and GIS (Table 2). The same patterns emerged in final sample sizes and relative efficiencies (Table 2) as were seen with the finer stratification (see Table 1). The smallest final sample size was with GBIS. GBIS also had the higher relative efficiency in comparison with GIS. Relative efficiencies were highest with smaller values of g and k. As with the results in table 1, relative efficiencies were improved with larger values of n_1 and smaller values of n_3. Overall, having the population area divided into six strata rather than 12 decreased the relative efficiency, with now the average being 1.91 for GBIS, and 1.61 for GIS (Table 2).

To confirm these trends, the simulations were repeated with 4 strata. Again, the largest final sample sizes among all set-up options were for CIS and then GIS (Table 3). Relative efficiencies patterns were the same as when the population was stratified into six and 12 strata, and overall were lower. The average relative efficiency was now 1.62 for GBIS, and 1.44 for GIS (Table 3).

4.2 Bladderpod Population

Our second example is from Morrison et al. (2008). The population shown in figure 2, is abundance scores of the Missouri bladderpod, *Lesquerella filiformis*, a small winter annual in the mustard family (*Brassicaceae*) (Rollins 1956; Rollins and Shaw 1973). The plant is considered threatened and is found only in a few locations in Missouri and Arkansas, USA.
The population shown is from an April 2003 survey in Bloody Hill Glade, at Wilsons Creek National Battlefield near Republic, Missouri. The site was divided into 5×5m quadrants, and shown in figure 2 is an abundance class ($1 = 1 - 9$ plants, $2 = 10 - 49$ plants, $3 = 50 - 99$ plants, $4 = 100 - 499$ plants) (Kelrick 2001). Following the methods of Morrison et al. (2008), we converted the abundance classes into counts for the 864 quadrants by generating a Poisson random deviate with expected value equal to the category midpoint.
The trends in the results from this population were consistent with the buttercup population. The relative efficiencies were generally highest when sampling was set up with small g or small k. Improvements in efficiencies were seen when the initial sample size, n_1, was larger, and when the sample effort within strata was controlled by a smaller value of n_3. For the designs other than with the smallest value of g and k, the relative efficiency of GBIS was higher with the smaller value of n_3. The average relative efficiency was 2.46 for GBIS and 1.95 for GIS (table 4).

5 Conclusion

We have proposed a modification for inverse sampling, called gap based inverse sampling. In this new design sampling continues until a gap is encountered. A gap can be the sequence of selected units that do not satisfy a condition (e.g., the condition can be the unit contains at least one rare event), following selection of a unit that does contain the rare event. The rules for sampling are as follows: sampling continues until a gap with length g is encountered; or until a pre-determined number of rare units, k, is found; or until n_3 units are selected.

Gap based sampling shows promise for field designs for rare events because it is an efficient design, and offers a way of limiting the final sample size. Here we illustrated the use in a stratified sampling context where the gap based design was used within each stratum. The gap based design, had the lowest sample size compared with GIS and CIS, and was the most efficient. The appeal of using stratified sampling with the gap based design for sampling rare events is that at times an entire stratum will have no rare events. Gap based sampling presents a mechanism for allowing the surveying team to move on to more productive strata.

Our simulation results give some insight into how to design an efficient survey. We found that having a small value of the pre-defined gap, g, and a small value for k, generally resulted in improved relative efficiencies.

The choice of a small value for g and for k, combined with stratified sampling can be a very effective way to control sampling effort. We recommend using small-sized strata with gap based sampling. With a rare population and small sized strata, many of the strata will be empty of the rare event. By using gap based sampling with small values for g and k, sampling effort will quickly move on from strata that contain no rare events. Sampling effort will be redirected to searching other strata. Setting the exact value of g, k, and the size of the strata will depend on the individual sampling situation, and the distribution of the population of interest, but these principles will apply in general. As with many other adaptive sampling designs it is recommended that the effect of different values of g, k, and other design characteristics are explored by simulation. Realistic population values may be available from previous surveys, or from an initial pilot survey.

With GBIS, we found that relative efficiency improved with larger values of n_1 and smaller values of n_3. This trend is the same as for the size of g and k. With a smaller value for n_3, survey effort is moved from unproductive strata on to searching other strata more quickly than if n_3 is large. Having a reasonably sized initial sample, n_1, improves relative efficiency, a trend seen in many other adaptive stratified designs.

We have presented GBIS with a broad class of stopping rules, that is, g, k and n_3. GBIS can have more simple rules, such as stopping when a gap is encountered without needing to count the number of rare units. Another alternative is to stop when either a gap is encountered, or the number of rare units, k, is found. Neither of these alternatives have an upper limit to the sample size, n_3, but for some situations their more simplistic form could be useful. We have provided estimators for these in appendix 2.

Appendix 1

For the Murthy estimator we have:

$$\hat{\mu}_y = \frac{1}{N} \sum_{i \in s} \frac{P(s \mid I_i = 1)}{P(s)} y_i$$

where I_i is an indicator function that takes 1 if the i^{th} unit is selected in the first selection. For simplicity we show event of "$I_i = 1$" by "i". $P(s)$ is the probability that s is chosen. $P(s|i)$ is the probability that s is chosen given that i is selected at the first place.
Then for unbiased estimators for the design we need only to calculate $\frac{P(s|i)}{P(s)}$. $P(i) = \frac{1}{N}$ and $\frac{P(s|i)}{P(s)} = \frac{P(s,i)}{P(i)P(s)}$, we need to calculate $\frac{P(s,i)}{P(s)}$ for the design.
We define $\frac{P(s,i)}{P(s)} = \frac{D_i}{D}$ where D_i is the number of ways that the selected sample s can be constructed such that unit i is the first selected unit, and D is the number of ways that the selected sample s can be constructed.
Because we set $g \geq n_1$, if $n = n_1$ then $D = n_1!$ and $D_i = (n_1 - 1)!$. For the case when $n = n_2$ (size of sample when we stop because of a gap) we have a situation as below (where the y's are the variables satisfying, and x's are the variables not satisfying, C):

$$\underbrace{x_1, x_2, y_3, x_4, y_5, ..., \underbrace{y_{n_2-g}}_{n_c^{th}-y}}_{n_c \text{ "}y\text{", and } n_{c'}-g \text{ "}x\text{"}}, \underbrace{x_{n_2-g+1}, ..., x_{n_2}}_{gap\ \ of\ \ size\ \ g}$$

In this situation the selected sample can be constructed if x's and y's are arranged in such way that there is no gap before the last y. We can simulate the situation by putting $n_{c'} - g$ cases in k places, where no place can tolerate more than $g - 1$ cases:

$$z_1 + z_2 + ... + z_{n_c} = n_{c'} - g; \tag{1}$$
$$z_l \in \mathbb{Z}; 0 \leq z_l < g; l = 1, 2, ..., n_c$$

Lemma 1 *Number of roots (number of response for vector $z = (z_1 + z_2 + ... + z_{n_c})$) of the equation 1 is*

$$\sum_{j=0}^{n_c} \binom{n_c}{j} (-1)^j \binom{n_{c'} - g + n_c - 1 - jg}{n_c - 1}$$

Proof 1 *We can solve the equation as*

$$\Upsilon\{z_1 + z_2 + ... + z_{n_c} = n_{c'} - g; 0 \leq z_l\} - \Upsilon\{z_1 + z_2 + ... + z_{n_c} = n_{c'} - g; \exists l, \;\; g \leq z_l\}$$

where Υ means "Number of roots".

$$\Upsilon\{z_1 + z_2 + ... + z_{n_c} = n_{c'} - g; 0 \leq z_l\} = \binom{n_{c'} - g + n_c - 1}{n_c - 1}$$

and

$$\Upsilon\{z_1 + z_2 + ... + z_{n_c} = n_{c'} - g; \textit{atleast for one } l, \;\; g \leq z_l\} = \binom{n_c}{1}\binom{n_{c'} - g + n_c - 1 - g}{n_c - 1} - \binom{n_c}{2}\binom{n_{c'} - g + n_c - 1 - 2g}{n_c - 1} + \binom{n_c}{3}(-1)^j\binom{n_{c'} - g + n_c - 1 - 3g}{n_c - 1} + ... + (-1)^n_c \binom{n_c}{n_c}(-1)^j\binom{n_{c'} - g + n_c - 1 - n_c g}{n_c - 1}$$

Then

$$D = n_c! n_{c'}! \sum_{j=0}^{n_c} \binom{n_c}{j} (-1)^j \binom{n_{c'} - g + n_c - 1 - jg}{n_c - 1}$$

For D_i the calculation is the same except because the first element is fixed we should find number of roots:

$$z_1 + z_2 + ... + z_{n_c - 1} = n_{c'} - g; \;\; 0 \leq z_l < g; l = 1, 2, ..., n_c - 1; \quad \textit{if the element is } y$$
$$z_1 + z_2 + ... + z_{n_c} = n_{c'} - g - 1; \;\; 0 \leq z_l < g; l = 1, 2, ..., n_c; \quad \textit{if the element is } x$$

At the end according to the calculation and above discussion we have

- for $\{y_i \in C, \;\; 1 < r < k\}$

$$\frac{P(s,i)}{P(s)} = \frac{(n_c - 1)! n_{c'}! \sum_{j=0}^{n_c - 1} \binom{n_c - 1}{j} (-1)^j \binom{n_{c'} - g + n_c - 2 - jg}{n_c - 2}}{n_c! n_{c'}! \sum_{j=0}^{n_c} \binom{n_c}{j} (-1)^j \binom{n_{c'} - g + n_c - 1 - jg}{n_c - 1}}$$

- for $\{y_i \in C', \;\; 1 < r < k\}$

$$\frac{P(s,i)}{P(s)} = \frac{n_c! (n_{c'} - 1)! \sum_{j=0}^{n_c} \binom{n_c}{j} (-1)^j \binom{n_{c'} - g + n_c - 2 - jg}{n_c - 1}}{n_c! n_{c'}! \sum_{j=0}^{n_c} \binom{n_c}{j} (-1)^j \binom{n_{c'} - g + n_c - 1 - jg}{n_c - 1}}$$

- for $\{y_i \in C, \;\; r = k > 1\}$

$$\frac{P(s,i)}{P(s)} = \frac{(n_c - 1)! n_{c'}! \sum_{j=0}^{n_c - 1} \binom{n_c - 1}{j} (-1)^j \binom{n_{c'} + n_c - 2 - jg}{n_c - 2}}{n_c! n_{c'}! \sum_{j=0}^{n_c} \binom{n_c}{j} (-1)^j \binom{n_{c'} + n_c - 1 - jg}{n_c - 1}}$$

- for $\{y_i \in C', \ \ r = k > 1\}$

$$\frac{P(s,i)}{P(s)} = \frac{n_c!(n_{c'}-1)!\sum_{j=0}^{n_c}\binom{n_c}{j}(-1)^j\binom{n_{c'}+n_c-2-jg}{n_c-1}}{n_c!n_{c'}!\sum_{j=0}^{n_c}\binom{n_c}{j}(-1)^j\binom{n_{c'}+n_c-1-jg}{n_c-1}}$$

- for $\{y_i \in C, 1 < r < k, n = N\}$ or $\{y_i \in C, 1 < r, g_n < g, n = n_3\}$

$$\frac{P(s,i)}{P(s)} = \frac{(n_c-1)!n_{c'}!\sum_{j=0}^{n_c}\binom{n_c}{j}(-1)^j\binom{n_{c'}+n_c-1-jg}{n_c-1}}{n_c!n_{c'}!\sum_{j=0}^{n_c+1}\binom{n_c+1}{j}(-1)^j\binom{n_{c'}+n_c-jg}{n_c}}$$

- for $\{y_i \in C', 1 < r < k, n = N\}$ or $\{y_i \in C', 1 < r, g_n < g, n = n_3\}$

$$\frac{P(s,i)}{P(s)} = \frac{n_c!(n_{c'}-1)!\sum_{j=0}^{n_c+1}\binom{n_c+1}{j}(-1)^j\binom{n_{c'}+n_c-1-jg}{n_c}}{n_c!n_{c'}!\sum_{j=0}^{n_c+1}\binom{n_c+1}{j}(-1)^j\binom{n_{c'}+n_c-jg}{n_c}}$$

- for $\{y_i \in C, r = 1, i = 1\}$

$$\frac{P(s,i)}{P(s)} = 1$$

- for $\{y_i \in C, r = 1, y_1 \in C'\}$ or $\{y_i \in C', r = 1, y_1 \in C\}$

$$\frac{P(s,i)}{P(s)} = 0$$

- for $\{y_i \in C', r = 1, y_1 \in C'\}$

$$\frac{P(s,i)}{P(s)} = \frac{1}{n_{c'}}$$

- for $\{r = 0\}$ or $\{n = n_1\}$

$$\frac{P(s,i)}{P(s)} = \frac{1}{n}$$

Appendix 2

- With a simple stopping rule based on g only the estimator is as follows:

$$\hat{\mu}_y = \begin{cases} \frac{\Phi(1,g)}{\Phi(0,g)}\bar{y}_{sc} + \frac{\Phi(0,g+1)}{\Phi(0,g)}\bar{y}_{sc'} & r > 1 \\ y_1 & r = 1, y_1 \in C \\ \bar{y}_{sc'} & r = 1, y_1 \in C' \\ \bar{y}_s & r = 0 \end{cases}$$

- With a stopping rule based on g and k, the estimator is as follows:

$$\hat{\mu}_y = \begin{cases} \frac{\Phi(1,g)}{\Phi(0,g)}\bar{y}_{sc} + \frac{\Phi(0,g+1)}{\Phi(0,g)}\bar{y}_{sc'} & 1 < r < k \\ \\ \frac{\Phi(1,0)}{\Phi(0,0)}\bar{y}_{sc} + \frac{\Phi(0,1)}{\Phi(0,0)}\bar{y}_{sc'} & r = k > 1 \\ \\ \frac{\Phi(0,0)}{\Phi(-1,0)}\bar{y}_{sc} + \frac{\Phi(-1,1)}{\Phi(-1,0)}\bar{y}_{sc'} & r < k, n = N \\ \\ y_1 & r = 1, y_1 \in C \\ \\ \bar{y}_{sc'} & r = 1, y_1 \in C' \\ \\ \bar{y}_s & r = 0 \end{cases}$$

References

[1] Brown, J.A., 2010. *Adaptive sampling of ecological populations.* In: Yue Rong (ed) Environmental Statistics and Data Analysis. ILM Publications , Hertfordshire, UK.

[2] Chistman M.C. and Lan F., 2001. Inverse adaptive cluster sampling, *Biometrics*, 57: 1096–1105.

[3] Finney D.J., 1949. On a method of estimating frequencies, *Biometrika*, 36(1): 223-234.

[4] Haldane J.B.S., 1945. On a method of estimating frequencies, *Biometrika*, 33: 222-225.

[5] Kelrick M.I., 2001. Missouri bladderpod monitoring protocol for Wilsons Creek National Battlefield. *In: U.S. Geological Survey, Northern Prairie Wildlife Research Center. Jamestown*

[6] Morrison L.W., Smith D.R., Young C.C. and Nichols D.W., 2008. Evaluating sampling designs by computer simulation: a case study with the Missouri bladderpod. *Population Ecology*, 50: 417-425.

[7] Rollins R.C., 1956. On the identity of Lesquerella angustifolia. *Rhodora*, 58: 199-202.

[8] Rollins R.C. and Shaw E.A., 1973. *The genus Lesquerella (Cruciferae) in North America.* Harvard University Press, Cambridge

[9] Salehi M.M. and Seber G.A.F.,2001. A new proof of Murthys estimator with applies to sequential sampling, *Aus. NZ. Stat.*, 43(3): 281–286.

[10] Salehi M.M. and Seber G.A.F.,2004. A general Inverse Sampling Scheme and its Application to Adaptive Cluster Sampling, *Aus. NZ. Stat.*, 46: 483-494.

[11] Wang, Y-G., Chen, Z.H., and Liu, J., 2004. General ranked set sampling with cost considerations. *Biometrics*, 60: 556–561.

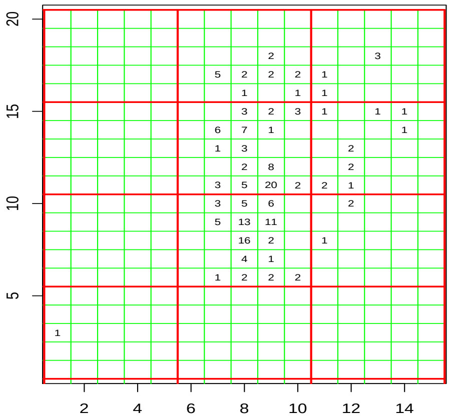

Figure 1: The sample population showing the counts within each quadrant of Castle Hill buttercups. There are $N = 300$ quadrants of size 10×10m. Here the population area is shown stratified in 12 stratum of size $N_h = 25$.

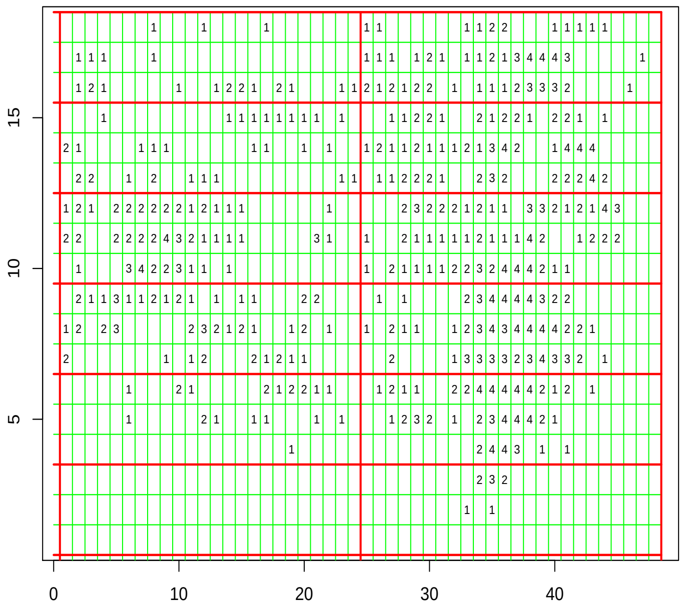

Figure 2: The bladderpod sample population, showing the abundance class of the counts of plants in each quadrant. There are $N = 864$ quadrants of size 5×5m. The population is stratified in 12 stratum of size $N_h = 72$.

Table 1: Expected final sample size and relative efficiency (RE), based on the Monte Carlo simulation and for stratified sampling with $N_h = 25$, of the buttercups for conventional Inverse Sampling (CIS); General Inverse Sampling (GIS); Gap Based Inverse Sampling (GBIS). The set-up values used in these designs, shown as within-stratum values, are: the length of the gap, g; the number of rare events, k; the initial sample size, n_1; and the sample size beyond which sampling stops, n_3.

n_3	n_1	k	g	$E(n_{CIS})$	$E(n_{\text{GIS}})$	$E(n_{\text{GBIS}})$	RE_{GIS}	RE_{GBIS}
13	6	2	8	200	128	94	4.16	5.09
			10	200	128	109	4.29	4.59
		4	8	238	141	107	1.90	2.10
			10	238	141	122	1.95	1.88
		6	8	258	150	116	1.68	1.61
			10	258	150	131	1.67	1.44
	8	2	8	200	176	95	5.80	6.69
			10	200	176	111	5.75	6.00
		4	8	238	199	111	2.42	2.58
			10	238	199	132	2.44	2.30
		6	8	258	216	128	1.75	1.66
			10	258	216	148	1.75	1.52
20	6	2	8	199	135	101	3.02	5.19
			10	200	135	116	2.96	4.40
		4	8	238	143	109	1.36	2.12
			10	238	143	124	1.35	1.84
		6	8	258	151	116	1.19	1.58
			10	258	151	132	1.18	1.43
	8	2	8	200	183	102	4.37	6.93
			10	200	183	118	4.24	6.03
		4	8	238	202	114	1.78	2.68
			10	238	202	134	1.74	2.31
		6	8	258	217	128	1.21	1.63
			10	258	217	149	1.23	1.49
		Average		232	170	119	2.55	3.13

Table 2: Expected final sample size and relative efficiency for stratified sampling with $N_h = 50$ for the buttercups.

n_3	n_1	k	g	$E(n_{CIS})$	$E(n_{\mathrm{GIS}})$	$E(n_{\mathrm{GBIS}})$	RE_{GIS}	RE_{GBIS}
30	10	4	15	190	132	94	1.94	2.47
			20	190	132	110	1.91	2.18
		6	15	210	149	111	1.44	1.58
			20	210	149	128	1.44	1.52
		8	15	230	164	125	1.37	1.37
			20	230	164	142	1.38	1.40
	15	4	15	190	140	102	2.65	3.11
			20	190	140	118	2.69	2.89
		6	15	210	151	113	1.73	1.92
			20	210	151	130	1.70	1.77
		8	15	230	164	126	1.43	1.46
			20	230	164	143	1.42	1.49
40	10	4	15	190	162	95	1.54	2.37
			20	190	162	115	1.53	2.07
		6	15	210	180	112	1.20	1.70
			20	210	180	133	1.21	1.52
		8	15	230	199	131	1.15	1.34
			20	230	200	152	1.15	1.32
	15	4	15	190	170	102	2.23	3.19
			20	190	170	122	2.27	2.91
		6	15	210	182	114	1.42	1.79
			20	210	182	135	1.45	1.78
		8	15	230	200	131	1.19	1.43
			20	230	200	152	1.20	1.35
Average				210	166	122	1.61	1.91

Table 3: Expected final sample size and relative efficiency for stratified sampling with $N_h = 100$ for the buttercups.

n_3	n_1	k	g	$E(n_{CIS})$	$E(n_{\mathrm{GIS}})$	$E(n_{\mathrm{GBIS}})$	RE_{GIS}	RE_{GBIS}
50	15	10	20	200	128	95	1.55	1.90
			30	200	128	112	1.56	1.70
		14	20	239	139	107	1.70	1.84
			30	239	139	123	1.69	1.79
		18	20	251	148	115	1.50	1.50
			30	250	148	131	1.50	1.53
	20	10	20	200	128	96	1.60	1.92
			30	200	128	112	1.60	1.75
		14	20	239	139	107	1.69	1.83
			30	239	139	123	1.69	1.79
		18	20	251	148	115	1.50	1.41
			30	250	148	131	1.51	1.50
70	15	10	20	200	165	106	1.21	1.71
			30	200	165	128	1.22	1.49
		14	20	239	179	121	1.33	1.61
			30	239	179	143	1.32	1.51
		18	20	250	190	131	1.30	1.38
			30	251	191	154	1.30	1.41
	20	10	20	200	165	106	1.24	1.70
			30	200	165	128	1.25	1.54
		14	20	239	179	121	1.33	1.70
			30	239	179	143	1.33	1.53
		18	20	250	190	132	1.30	1.41
			30	250	190	154	1.30	1.42
Average				230	158	122	1.44	1.62

Table 4: The expected final sample size, and relative efficiency for stratified sampling with $N_h = 72$ for the bladderpods.

n_3	n_1	k	g	$E(n_{CIS})$	$E(n_{\text{GIS}})$	$E(n_{\text{GBIS}})$	RE_{GIS}	RE_{GBIS}
25	4	2	6	142	97	62	2.37	3.16
			8	141	97	68	2.34	3.03
		3	6	176	118	75	1.68	2.11
			8	176	118	84	1.67	2.04
		4	6	211	140	92	1.51	1.69
			8	211	140	103	1.51	1.72
	6	2	6	142	111	76	3.04	3.49
			8	141	111	82	3.06	3.61
		3	6	176	126	83	2.12	2.44
			8	176	126	92	2.12	2.40
		4	6	211	143	95	1.70	1.84
			8	211	143	106	1.70	1.86
35	4	2	6	141	110	61	2.11	3.21
			8	142	111	68	2.05	3.05
		3	6	176	134	75	1.46	2.11
			8	176	134	84	1.47	2.08
		4	6	211	159	92	1.32	1.72
			8	211	159	103	1.33	1.70
	6	2	6	141	124	76	2.81	3.64
			8	142	124	82	2.71	3.61
		3	6	176	142	83	1.86	2.35
			8	176	143	92	1.88	2.49
		4	6	211	162	95	1.47	1.79
			8	211	162	106	1.48	1.84
Average				176	131	85	1.95	2.46